

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

METHOD AND SYSTEM FOR GENERATING DATA PACKET IN DIFFERENT KINDS OF NETWORK

Patent number: JP11112574
Publication date: 1999-04-23
Inventor: MULLIGAN GEOFFREY
Applicant: SUN MICROSYST INC
Classification:
- international: H04L12/56; H04L12/46; H04L12/28; H04L29/06
- european:
Application number: JP19980175701 19980623
Priority number(s):

Also published as:

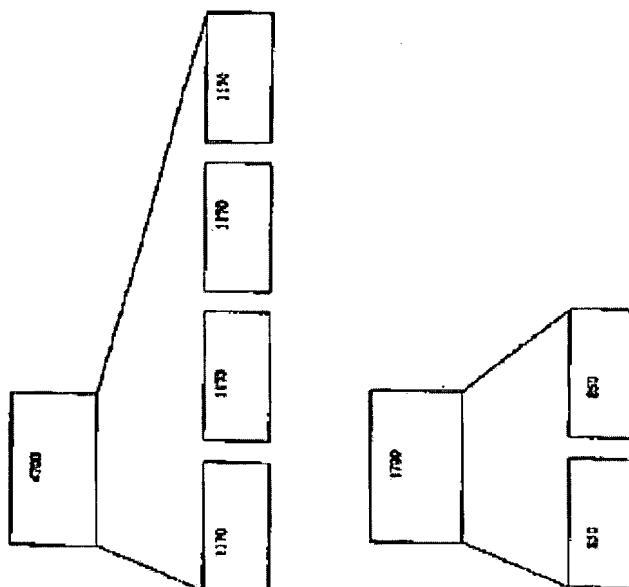


US6212190 (B1)
JP11112574 (A)

Abstract of JP11112574

PROBLEM TO BE SOLVED: To reduce quantity for fragmenting datagram into segments and to extend valid network band width associated with packet transmission on a network by comparing the size of a maximum transmission unit(MTU) that can be transmitted at a prescribed route with the sizes of transmitted packets and processing and transmitting the packets when the sizes of the packets are large.

SOLUTION: When the datagram segment is larger than MTU of a reception side network, a router fragments the datagram into plural smaller datagrams before the datagram is transmitted. The datagram of 4700 bytes is distributed to four datagrams provided with data of 1170 bytes and the datagram of 1700 bytes is divided into two smaller datagrams of 850 bytes for the network of 1500 bytes by MTU, for example. Data is uniformly distributed among the datagrams, the maximum fragmented size is reduced and the probability of fragmentation can be reduced.



Data supplied from the *esp@cenet* database - Worldwide

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-112574

(43) 公開日 平成11年(1999) 4月23日

(51) Int.Cl.⁶ 識別記号

H 0 4 L 12/56
12/46
12/28
29/06

F I

H 0 4 L 11/20 1 0 2 A
11/00 3 1 0 C
13/00 3 0 5 B

審査請求 未請求 請求項の数 8 O L (全 22 頁)

(21) 出願番号 特願平10-175701

(22) 出願日 平成10年(1998) 6月23日

(31) 優先権主張番号 08/880200

(32) 優先日 1997年 6月23日

(33) 優先権主張国 米国 (US)

(71) 出願人 591064003

サン・マイクロシステムズ・インコーポレ
ーテッド

SUN MICROSYSTEMS, IN
CORPORATED

アメリカ合衆国 94303 カリフォルニア
州・バロ アルト・サン アントニオ ロ
ード・901

(72) 発明者 ジェフリー・マリガン

アメリカ合衆国・80920・コロラド州・コ
ロラド スプリングス・クローバーデイル
ドライブ・2175

(74) 代理人 弁理士 山川 政樹

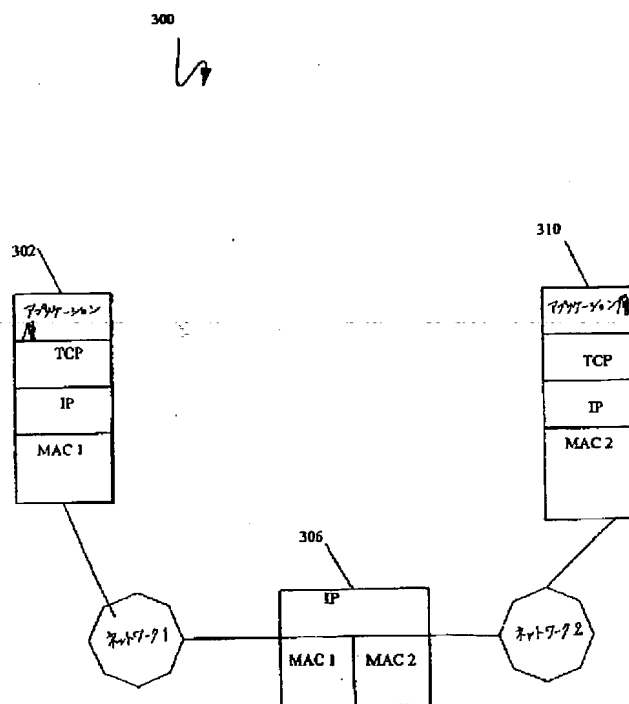
最終頁に続く

(54) 【発明の名称】 異種ネットワークでデータ・パケットを生成する方法およびシステム

(57) 【要約】

【課題】 ネットワーク上の異なるルートで送信されるパケットをより早く送ることができるように生成する改良された方法およびシステムを提供する。

【解決手段】 所定のルートで送信可能な最大送信単位 (MTU) を決定する。次に、ネットワーク上で送信する各パケットのサイズがMTUのサイズと比較される。この比較でパケットがMTUより大きいとわかると、パケットはさらに処理されてからルート上で送信される。その処理の特徴は従来のように最大MTUで分割するのではなく、個々のパケットのサイズ送ることができるサイズで、しかも均等のMTUになるように分割する。



【特許請求の範囲】

【請求項1】 それぞれのルートが異なるサイズの packets を伝送でき、packets のサイズが packets 内に含まれる記憶単位の総数によって決まる、ネットワーク上の異なるルートで送信される packets のサイズを決定するためのコンピュータで実施する方法であって、ネットワーク上で送信される packets 内の記憶単位の総数をネットワーク上の所与のルートに関連する最大伝送単位 (MTU) と比較するステップと、比較の結果が packets 内の記憶単位の総数が MTU より大きいことを示している場合に packets をさらに、packets 内の記憶単位数を MTU 記憶単位を保持できる1つまたは複数の packets 内に記憶された実質的に均等な単位のグループに分離することによって処理するステップとを含む方法。

【請求項2】 所定のルート上で送信可能な最大伝送単位 (MTU) を受信するステップをさらに含む請求項1に記載の方法。

【請求項3】 それぞれのルートが異なるサイズの packets を伝送でき、packets のサイズが packets 内に含まれる記憶単位の総数によって決まる、ネットワーク上の異なるルートで送信される packets のサイズを決定するように構成された装置であって、ネットワーク上で送信される packets 内の記憶単位の総数をネットワーク上の所与のルートに関連する最大伝送単位 (MTU) と比較するように構成された機構と、比較の結果が packets 内の記憶単位の総数が所与のルートに関連する MTU より大きいことを示している場合に packets をさらに処理するように構成された機構と、packets 内の記憶単位数を MTU 記憶単位を保持できる1つまたは複数の packets 内に記憶された実質的に均等な単位のグループに分離するように構成された機構とを含む装置。

【請求項4】 所定のルート上で送信可能な最大伝送単位 (MTU) を受信するように構成された機構をさらに含む請求項3に記載の装置。

【請求項5】 それぞれのルートが異なるサイズの packets を伝送できるネットワーク上の異なるルートで送信される packets を生成できるコンピュータ可読コードを記録した記録媒体であって、前記コードが、ネットワーク上で送信される packets 内の記憶単位の総数をネットワーク上の所与のルートに関連する最大伝送単位 (MTU) と比較するように構成された第1のコード部分と、比較の結果が packets 内の記憶単位の総数が MTU より大きいことを示している場合に packets をさらに処理する第2のコード部分と、packets 内の記憶単位数を MTU 記憶単位を保持できる1つまたは複数の packets 内に記憶された実質的に均等な単位のグループに分離するように構成された第3のコード部分とを含むことを特徴とするコンピュータ・プログラムを記録した記録媒体。

【請求項6】 所定のルート上で送信可能な最大伝送単位 (MTU) を受信するように構成された第8のコード部分をさらに含む請求項5に記載の記録媒体。

【請求項7】 搬送波内で実施され、プロセッサによって実行されると、それぞれのルートが異なるサイズの packets を伝送でき、packets のサイズが packets 内に含まれる記憶単位の総数によって決まる、ネットワーク上の異なるルートで送信される packets のサイズを、ネットワーク上で送信される packets 内の記憶単位の総数をネットワーク上の所与のルートに関連する最大伝送単位 (MTU) と比較するステップと、比較の結果が packets 内の記憶単位の総数が MTU より大きいことを示している場合、さらに、packets 内の記憶単位数を MTU 記憶単位を保持できる1つまたは複数の packets 内に記憶された実質的に均等な単位のグループに分離することによって packets を処理するステップとを実行することによって決定する命令シーケンスを表すコンピュータ・データ信号。

【請求項8】 所定のルート上で送信可能な最大伝送単位 (MTU) を受信するステップをさらに含む請求項7に記載の搬送波内に具体化されたコンピュータ・データ信号。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、一般にネットワーク通信、より詳細に言えば、コンピュータの異種ネットワーク上でデータ・packets を生成する方法およびシステムに関する。

【0002】

【従来の技術】ネットワーク・コンピューティングは最近10年間で著しい速度で発展した。ネットワーク・コンピューティング環境では、ユーザはネットワークに接続された複数のコンピュータにアクセスできる。ネットワーク・コンピューティングのトップ企業である Sun Microsystems Inc. は、コンピューティング市場におけるこの発展する分野の商業的成功を強調すべく「The Network Is the Computer」TM というスローガンを中心に市場キャンペーンを張っているほどである (The Network Is the Computer, Sun, Sun のロゴ、Sun Microsystems, Solaris, Ultra, Java は米国およびその他の諸国の Sun Microsystems Inc. の商標または登録商標。すべての SPARC 商標は、米国およびその他諸国の SPARC International, Inc. のライセンスの下で使用され、その商標または登録商標。SPARC 商標を付けた製品は Sun Microsystems Inc. が開発した

アーキテクチャに基づいている。UNIXは米国およびその他の諸国の登録商標であり、X/Open Company, Ltd. を通じて排他的にライセンスされる。)。最近、数百万人のユーザがインターネットおよびワールド・ワイド・ウェブ上で利用できる数千のコンピュータにあるコンピューティング・リソースを開発するようになって、このスローガンは現実のものとなった。

【0003】インターネットおよびワールド・ワイド・ウェブの普及は標準のネットワーク・プロトコルと異なるネットワークを結合するルータを使用したことによる。TCP/IPなどの通常のネットワーク・プロトコルはアプリケーション層、処理層、ホスト間プロトコル層(TCP/UDP)、インターネット・プロトコル層(IP)、ネットワーク・プロトコル層および物理層を含む。しばしばルータを使って物理層とネットワーク層にある異なるネットワークからの情報を変換している。特に、異なる物理層は、物理メディアを介して送信できる最大伝送単位(MTU)を特に規定している該当するメディア・アクセス層(MAC)を備える。一般に、MTUはすべてのネットワーク・セグメントが送信可能な最大送信単位の数として定義される。変換処理の一環として、ルータはMTUが大きいネットワークから送信されるパケットをMTUがより小さいネットワーク上のより小さいセグメントに断片化することを要求される。

【0004】例えば、インターネットは異なる物理メディアと、イーサネット、IEEE 802.3 (CSMA/CD)、IEEE 802.4 (トークン・リング)、PPP/SLIP (ダイヤルアップ)、T-1、T-3、およびCDMA、TDMA、AMPS、GSMなどの無線技術の空中インタフェースを含むそれに対応するMAC層を結合する異種IPネットワークである。上位層ではこれらの異なるネットワークはTCP、UDPなどのホスト間プロトコルおよびメール、フィンガ、およびftpなどのアプリケーション・プロトコルを共用する。異なるネットワーク間ですらこれらの上位層プロトコルは互換性がある。ただし、ネットワークおよびMAC層でこれらの異なるネットワークのMTU値(1990年11月現在、インターネット上の様々なデータ・リンクに関連する一般的MTU(バイト単位)には、以下のものがある。: 65535 Theoretical Maximum (RFC 791); 65535 Hyperchannel (RFC 1044); 17914 IBM Token Ring (RFC 791); 8166 IEEE 802.4 (RFC 1042); 4464 IEEE 802.5 (最高4Mb) (RFC 1042); 4352 FDDI (改訂版) (RFC 1188); 2048-4352 Wideband Network (RFC 1188); 2002 IEEE 802.5 (4Mb推奨) (RFC

1042); 1536-2002 Exp. EthernetNet (RFC 895); 1500 Ethernet Network (RFC 894); 1500 Point-to-Point (デフォルト) (RFC 1134); 1492 IEEE 802.3 (RFC 1042); 1006-1492 SLIP (RFC 1055); 1006 ARPANET (BBN 1822); 576-1006 X.25 Network (RFC 877); 544 DEC IP Portal; 512 Netbios (RFC 1088); 508 IEEE 802/Source-Rt Bridge; (RFC 1042); 508 ARCNET (RFC 1051); 296-508 Point-to-Point (低遅延) (RFC 1144); 68 Official minimum MTU. RFC番号を使って、インターネット・エンジニアリング・タスク・フォース(IETF)発行の対応する文書を見つけることができる。)は多種多様で、サイズの互換性がないことが多いデータグラム・セグメントを生成する。

【0005】データグラムは接続レスのネットワーク上で送信される情報パケットであるため、MTUの非互換性の問題が生じる。換言すれば、その中で情報パケットの移動が制限される接続が送信元ノードと宛先ノードの間に存在しない。その代わりに、各パケットは通常異種ネットワークの異なるパス上で送信され、宛先ノードに到着する。その結果、各中継ネットワークのMTUに合わせて各データグラムを異なる量で断片化することができる。例えば、イーサネット・ネットワークから送信されるデータグラム・セグメントは通常1500バイト長であり、より小さいいくつかのデータグラム・セグメントに断片化してX.25 (MTU 576) または Netbios (MTU 512) 物理ネットワーク層で送信される必要がある。このように、インターネットで用いられるルータはより小さなデータグラム・セグメント・サイズを必要とする、またはこれらのセグメントをネットワークの他の部分で互換性があるより小さいセグメントに断片化する大量のプロセッサ・サイクルを費やす。

【0006】残念ながら、連続的な断片化によってネットワークの相互運用性を達成するとデータ処理および送信オーバーヘッドが増加するためにネットワークのパフォーマンスは低下する。当初、大きいセグメントがますます小さいセグメントに縮小されてMTUがますます小さくなる物理ネットワーク層を備えたネットワーク上で送信されるにつれて、断片化がルータの処理能力を消耗させる。受信側では、宛先ホストはより多くの処理能力とリソースを使って拡張されたセグメント・プールをバッファに入れてその拡張されたセグメント・プールを元のメッセージに組み直す。さらに、ネットワーク上で追加のパケットを送信すると送信されるパケットのヘッダ情

報の量が増え、有効なネットワーク帯域幅が減少する。

【0007】断片化の制御は接続レス・ネットワーク上では特に困難である。それは、前述したように、各セグメントが異なるルートで送信できるからである。接続レス・ネットワーク上のデータグラムとして知られるいくつかのセグメントは大きいMTUを備えたルート上を送信でき、断片化の必要はない。しかし、ネットワークのあるセグメントが障害になるか不正確に組立てられると、ネットワーク・トポロジが変更され、以降のセグメントはMTUがより小さくまた広範囲の断片化を必要とするルート上で送信できる。複雑なルーティングによって送信元ホストは断片化を最小にする適切なデータグラム・サイズを選択することが困難になる。例えば、インターネット上で使用される一次ルートが8166バイトのMTUを備えたIEEE802.4ネットワークを含み、二次ルートが576バイトというはるかに小さいMTUを備えたX.25ネットワークを含むと仮定する。送信パケット数を最小化するため各データグラムが2つのルートの大きい方のMTUに基づき、8166バイトを含むと仮定する。当初、一次ルート上で送信されるデータグラムは断片化を必要としない。しかし、一次ルートが使用できなくなるか話中の場合、ネットワーク・トポロジが変更されて576バイトというより小さいMTUを備えた二次ルートがトラフィックを引き受ける。したがって、ルータは残りのデータグラムのそれぞれを断片化し、上述したように有効なネットワーク帯域幅を縮小するインターネット上で送信する。接続レス・ネットワーク上で断片化を最小にするデータグラム・サイズを選択するのはまさしく困難な作業である。

【0008】断片化全体を減らす一方でネットワーク帯域幅を最適化するデータグラム・サイズを選択する現在の技法は、まだ十分に成功していない。一般にはデータはローカル・ネットワーク上で送信されローカル・ネットワークMTUに基づいてデータグラムを生成する。残念ながらこの技法では、ローカル・ネットワークのMTUサイズより小さいMTUを備えたリモート・ネットワーク上でデータグラムが送信される場合は断片化が実行され、オーバーヘッドが増加する。例えば、データがルータ装置とX.25ネットワークで結合されたあるイーサネットから別のイーサネットへ送信されると仮定する。さらにイーサネット・ネットワークのMTUが1500バイト、2つのイーサネット・ネットワーク間のX.25ネットワークのMTUが576バイトと仮定する。すると、この第1の技法を使って生成されたデータグラムは、X.25ネットワークをまたがる2つの576バイトのデータグラムと1つの348バイトのデータグラムに断片化される。この技法は通常データグラムがローカル・ネットワークにある限り断片化を減少させない。

【0009】別の技法は、送信元ホストとして同じローカル・ネットワークにもサブネットにも接続されてい

ないネットワークに送信する場合に、小さい方の576バイトすなわちファーストホップMTUに基づいてデータグラムの大きさを決定する。この場合、あるパケットが通過したルータ、スイッチ、またはブリッジの数を数えることでホップ・カウントがネットワーク上の移動距離を表す。あるパケットに関連付けられたホップ・カウントMはパケットがおよそM個の異なるネットワークを通過したことを意味する。当然の結果として、ネットワークのファーストホップMTUは送信側ネットワークからちょうど1ホップ分離されている。残念なことに、MTUを選択するこの技法は通常必要な数に足りないデータグラムを生成しネットワーク機能が低下して使用可能なネットワーク帯域幅を浪費する。さらに、データグラムが576未満のMTUを備えたネットワークを通過すると断片化が行われる。

【0010】第3の技法はネットワーク上の送信元ノードと宛先ノードの間のルートのパスMTU(PMTU)に基づいてデータグラムのサイズを決定する。IPネットワークに適用されるPMTU検出技法は、Jeffrey MogulとSteve Deering著の、インターネット特別技術調査委員会(IETF)に関連して出版された「Path MTU Discovery」という題のRFC1191に記載されている。要するに、この技法はIPヘッダの「Don't Fragment」(DF)ビットを使って任意のネットワーク・パスのPMTUを動的に発見する。最初、送信元ホストはPMTUをファーストホップMTUと仮定してDFビットを立ててそのパス上で全データグラムを送信する。これらのデータグラムが任意のネットワーク・パスで何らかのルータによって断片化なしに転送するには大きすぎる場合、当該ルータはサイズ超過のデータグラムをドロップし送信元ノードに「断片化が必要でDFビットはオン」を示す特別のコードを付けたインターネット制御メッセージ・プロトコル(ICMP)デスティネーション・アンリーチャブル(Destination Unreachable)メッセージを返送する。これに回答して、送信元ノードはデータグラムが断片化なしに送信できるまでその大きさをますます小さくして上記の処理を繰り返す。残念なことに、インターネットなどの接続レス・ネットワーク上で送信されるそれぞれの連続的なデータグラムは別のルートを通るため、同じネットワーク上で送信された以前のデータグラムとは異なるPMTU値を備える。このように、Path MTU Discoveryを用いたシステムでさえ断片化を減らすことができず、したがってパケット送信量が増えるとネットワークの有効帯域幅が減ることになる。

【0011】

【発明が解決しようとする課題】したがって、必要なものは断片化の量を削減し、ネットワーク上でのパケット送信に関連付けられた有効ネットワーク帯域幅を拡張す

るネットワークで用いるパケット生成の方法およびシステムである。

【0012】

【課題を解決するための手段】本発明はネットワーク上の異なるルートで送信するパケットを生成する改良された方法およびシステムを提供する。インターネットなどの巨大なネットワークでは、ネットワーク上の各ルートはより小さい相互運用が可能なセグメントに断片化されるまではネットワーク間で直ちには送信できない異なるサイズのパケットを送信する。まず、この技法は所定のルートで送信可能な最大送信単位 (MTU) を決定する。次に、ネットワーク上で送信する各パケットのサイズがMTUのサイズと比較される。この比較でパケットが所定のMTUより大きいとわかると、パケットはさらに処理されてからルート上で送信される。さらに別の処理ではまずパケットに含まれる全送信単位数をMTU値で割る。この除算の結果はDCOUNT記憶ユニットに一時記憶され、除算の余りはRCOUNT記憶ユニットに記憶される。RCOUNTの余りが非ゼロの場合、DCOUNTの値が1つ増分される。DCOUNT値は現在の技法を用いてパケットを送信する場合の最小データグラム数を示す。次に、元のパケットに含まれる送信単位はDCOUNTパケットに公平に分配され、ネットワーク・ルートでの送信用に準備される。

【0013】本発明は従来技術では使用できなかったいくつかの利点がある。まず、本発明で提供される技法は、送信されるデータグラム数が実質的に変更されないためネットワーク・オーバーヘッド量を増やさずに断片化の確率を減らす。最良のケースではこれらの技法は大幅に断片化を減らして有効ネットワーク帯域幅を拡張するが、最悪のケースではこれらの技法の有効性は他の技法と同じ程度である。

【0014】本発明はネットワーク・プロトコルに非互換性を持ち込まずにネットワーク・パフォーマンスを向上させるため、また有利である。大半の大きいネットワークではネットワーク・プロトコル内の低レベル層の変更は不可避免的に非互換となりネットワークの相互運用性が減少する。これはIPやTCP/IPなどのプロトコルを実行する数百、数千のコンピュータが直ちに變更できないインターネットについて当てはまる。本明細書に記載したネットワーク処理の改良は基本的にIPプロトコル動作を變更しないため、本発明を使用してもこのジレンマは発生しない。したがって、本発明の教示によってネットワークの相互運用性を維持し、データ・パケットを送信するために有効なネットワーク帯域幅を改善するという困難な問題は明快に解決される。

【0015】

【発明の実施の形態】図1に本発明の一実施形態を実施するコンピュータ・ネットワーク100を示す。この例では、送信元ノード102と宛先ノード114がルータ

106および110を使ってネットワーク104、108、および112を含む一次ルート上で相互にデータを送信する。また別に、送信元ノード102および宛先ノード114はルータ106および中間ノード116を使ってネットワーク104、118、および112を含む二次ルート上で通信を行うことができる。コンピュータ・ネットワーク (図1) は通常他のデータ装置へ1つまたは複数のネットワークを介してデータ通信を行う。例えば、ネットワーク100は送信元ノード102および宛先ノード114の間に一部現在「インターネット」という通称があるワールドワイドなパケット・データ通信ネットワークを使ってネットワーク接続を行うことができる。インターネットは電気、電磁気、および光信号を使ってさまざまなタイプの情報を表すデジタル・データ・ストリームを伝送する。コンピュータ・ネットワーク100を介して伝送される信号は送信元ノード102への、また送信元ノードからのデジタル・データを伝送し、情報を送信する搬送波の代表的な形式である。

【0016】送信元ノード102はネットワーク100を介してメッセージを送信し、プログラム・コードを含むデータを受信する。インターネットの例では、宛先ノード104はアプリケーション・プログラムの要求コードをネットワーク100を使ってインターネット経由で送信元ノード102に送信できる。本発明は、その種のダウンロードされたアプリケーションの1つが異種ネットワーク上でデータ・パケットを生成する方法とシステムであり、本明細書に記載される。受信されたコードは受信時に送信元ノード102によって実行され、または記憶装置に記憶されて後ほど実行される、あるいはその両方の処理がされる。このように、送信元ノード102は搬送波の形式でアプリケーション・コードを入手できる。

【0017】上記のルータおよびノードは汎用コンピュータの内部で実行される処理として、またはネットワーク上で送信されるデータ・パケットを受信して処理するための専用のスタンドアロン装置として実施できることは当業者は理解するであろう。これに依じて、これらの装置の一般のコンポーネントを図2に示す。

【0018】図2について説明する。同図には、ネットワーク100 (図1) 上のルータまたはノードとして動作するよう構成されたコンピュータ・システムの必須コンポーネントが示されている。図2のコンピュータ・システム202は第1のネットワーク・インタフェース212、第2のネットワーク・インタフェース211、プロセッサ214、一次記憶216、二次記憶218、および前記の要素間の通信を容易にする入出力インタフェース220を含む。ネットワーク・インタフェース212はコンピュータ・システム202をネットワーク100 (図1) に結合し、コンピュータ・システム202とネットワーク100 (図1) 上のその他のノード間の通

信を容易にする。ネットワーク・インタフェース211はコンピュータ・システム202を第2のネットワーク（図示せず）に結合し、コンピュータ・システム202とネットワーク100（図1）上のその他のノードまたは第2のネットワーク（図示せず）上のノード間の通信を容易にする。ルータ装置は通常、複数の異なるネットワークを結合するために、複数のネットワーク・インタフェースを備えていることは当業者には周知である。したがって、図2の代表的なコンピュータ・システム202は2つの異なるネットワーク用に構成されているが、1つまたは複数のネットワーク・インタフェースを追加することで別のネットワークに結合できる。

【0019】図2で、コンピュータ・システム202内のプロセッサ214は入出力インタフェース220を介して一次記憶216からのコンピュータ命令を取り出す。これらの命令を受信すると、プロセッサ214はこれらのコンピュータ命令を実行する。これらのコンピュータ命令を実行すると、プロセッサ214は一次記憶216のデータを検索し、記憶に書き込み、1つまたは複数のコンピュータ・ディスプレイ装置（図示せず）に情報を表示し、1つまたは複数の入力装置（図示せず）からコマンド信号を受信し、二次記憶218または送信元ノード102、宛先ノード114、中間ノード116、ルータ106、またはルータ110などのネットワーク100（図1）上のその他のノードのデータを受信しあるいはデータを書き込むことができる。プロセッサ214はまた、図2には図示していない。コンピュータ・システム202に結合された他のネットワークに関するノードに対して上記の機能を実施することができる。一次記憶216および二次記憶218は、ランダム・アクセス・メモリ（RAM）、読み出し専用メモリ（ROM）、磁気記憶装置およびCD-ROMなどの光記憶メディアを含めてあらゆる種類のコンピュータ記憶を無制限に含むことができることを当業者はまた理解する。プロセッサ214はカリフォルニア州Mountain ViewのSun Microsystems, Inc.製のSPARC互換プロセッサ、UltraSpark互換プロセッサまたはJava互換プロセッサのいずれかである。別に、プロセッサ214はカリフォルニア州CupertinoのApple, Inc.製のPowerPCプロセッサ、Intel CorporationまたはAMD、およびCyrilx製のすべてのPentiumまたはx86互換プロセッサ、または他のすべての特殊用途のプロセッサに基づくことができる。

【0020】図2について説明する。同図では、一次記憶216がコンピュータ・リソースを管理するオペレーティング・システム222を含む。このオペレーティング・システムは、Solarisオペレーティング・システムか、ネットワーク・コンピューティング環境でのデータ・パケットおよびデータグラムの処理に関連付け

られた処理要件をサポートする能力があるすべてのオペレーティング・システムであることが好ましい。別の実施形態では、オペレーティング・システム222はCisco Inc.のルータ装置に使われるインターワーキング・オペレーティング・システム（IOS）またはAscend Inc.、Bay Network, Inc.、または3Com, Inc.のルータ装置に使われる類似のオペレーティング・システムである。また一次記憶にはTCP/IP、X.25、SNAなどの1つまたは複数のプロトコル・スタック224、または異なるネットワーク間でデータグラムを変換するNetWare（NetWareは米国およびその他諸国のNovell, Inc.の登録商標）などのネットワーク・オペレーティング・システムの一部も含まれる。本発明の一実施形態ではこれらのプロトコル・スタック224と、データグラム生成に用いるデータグラム生成器226が連動している。

【0021】実際、図2に示すルータ202は異なるネットワーク間でデータグラムを変換するのに必要なプロトコル・スタックの一部しか含まない場合がある。例えば、IPネットワーク内ではルータは通常2つの異なるネットワークでは異なるIP層の下層を全て含む。図3について説明する。同図で、ルータ装置306は、第1のネットワーク上で使用される下位レイヤ・メディア・アクセス・プロトコル1（MAC1）と第2のネットワーク上で使用されるメディア・アクセス・プロトコル2（MAC2）に加えてIP層を含む。上述したように、異なるネットワーク・メディアはそれに対応する異なるMACプロトコルを利用する。一方、異なるネットワーク上の各ノードはネットワーク・アクセス層、トランスポート層、およびアプリケーション層を含むアプリケーションを処理するプロトコル・スタック一式を含む。これらの追加の層はエンド・ユーザ、アプリケーション、またはリモート装置に上位レベル機能を提供する。これに合わせて、図3では、ノード302および310がIP層および該当するメディア・アクセス・プロトコルに加えてアプリケーション層とTCP層を含むことが示される。

【0022】この時点で、異種コンピュータ・ネットワークでのIP層の役割の基本説明が本発明の舞台を設定する上で役立つであろう。図4AはIP層を使ってX.25パケット交換ネットワーク407上の第1のノード402からイーサネット・ローカル・エリア・ネットワーク412上の第2のノード414へデータを送信する方法を示すブロック図である。各プロトコル・スタックの管理と、送信されたデータの処理に関連付けられたハードウェアは、この例ではIPプロトコルの動作を強調するために省略している。図4Bは図4Aに示すIPネットワーク上で送信するデータ・パケットへの変更をステップ1～15に示す図である。図4Bの各セグメント

は簡潔に次の略号で示す。トランスポート・ヘッダ (T-H)、インターネット・プロトコル・ヘッダ (I-P-H)、X. 25 パケット・ヘッダ (P-H)、LAP-B リンク・ヘッダ (L-H)、LAB-B トレーラ (L-T)、LLC ヘッダ (LLC-H)、MAC ヘッダ (MAC-H)、MAC トレーラ (MAC-T)。

【0023】図4Aについて説明する。同図では、第2のノード414がノード402に返送するデータがTCP層に対応するIPヘッダおよびトランスポート・ヘッダを備えたデータグラムであるインターネット・プロトコル・データ単位にカプセル化される。IPネットワーク上でデータグラムを送信するため、データグラムIPヘッダは第1のノード402のアドレスを含み、下位レベル・ネットワーク・パケット・ヘッダはルータ410のMACアドレスを含む。ルータ410のMACアドレスは第1のノード402のアドレスと共に含まれるが、これは第1のノード402から直接第2のノード414にアクセスできないためである。ルータ410は、イーサネット・ネットワーク412およびX. 25 ネットワーク407で使われる、異なるネットワーク・プロトコルの間の連絡機能として動作する。この例では、X. 25 パケット交換ネットワーク層3を使ってIPパケットがX. 25 ネットワーク上で移動する際にカプセル化している。最終的に、カプセル化されたデータグラムはX. 25 ネットワーク407を介して第1のノード402に送信され、ここで元のIPデータグラムが使用できるようになり処理が可能となる。ここではX. 25 ネットワーク動作の詳細は本明細書の範囲外であるが、X. 25 ネットワークによって元のIPデータグラムが変更されないことに注意するべきである。

【0024】実際にIPデータグラムを送信する前に、ルータ410はIPデータグラムがデータグラムを受信するネットワークのMTUを超えないことを確認する必要がある。例えば、データグラム・セグメントが受信側ネットワークのMTUより大きい場合、ルータ410はそのデータグラムを1つまたは複数のより小さいデータグラムに断片化する必要がある。図4Aで、X. 25 上で送信されるMTUが576のデータグラムはMTUが1500バイトのイーサネット・ネットワーク (IEEE 802.3) から受信した場合、少なくとも3つのデータグラムに分割する必要がある。上述したように、データグラムが小さくなると送信されるヘッダ情報の割合が増え、通常はパケットの再送数が増えるため、断片化によってネットワークの効率は低下する。

【0025】過去において、断片化を抑えるための技法は適切なデータグラムのサイズに焦点を当てていたが、実際のデータグラム内にデータがどのように分配されるかを見過ごしていた。最も一般的な技法はデータグラムを順次データで満たし、結果的に一連の大小のデータグラムが出来上がる。図5Aおよび図5Bに、従来技術で

MTUが1500バイトのネットワークで4700バイトのデータグラムと1700バイトのデータグラムを複数のデータグラムに不均等に断片化する処理を示す。この方法では、図5Aの4700バイトのデータグラムはより小さい1500バイトのデータグラム3つと4番目の200バイトのデータグラムに分配され、図5Bの1700バイトのデータグラムはより小さい1500バイトのデータグラム2つと200バイトのデータグラム1つに分配される。データグラムが比較的大きいとMTUがそれより小さいネットワークに遭遇することが多いため、従来技術では頻繁な断片化は珍しくなかった。

【0026】これとは対照的に、本発明の実施形態はこれらのデータグラム間に均等にデータを分配し、最大断片化サイズを減らして断片化の確率を減らす技法を提供することである。IPデータグラムを生成するこの新しい技法によるデータの均等な分配の例が図5Cおよび図5Dである。この方法では、本発明の原理によって分割された4700バイトのデータグラムはそれぞれが1170バイトのデータを備えた4つのデータグラムに分配され、図5Dの1700バイトのデータグラムはより小さい850バイトのデータグラム2つに分配される。この明快な解決法によって、ネットワークの非互換性を持ち込まずにデータグラム生成に変更を加えることでネットワーク効率は向上する。その結果、断片化されずにネットワークを通過できるデータグラムの数は増える。以下に、IPネットワークに関して本発明の一実施形態によるデータグラム生成を詳述する。

【0027】まず、本発明はネットワークのMTUに決定技法があることを前提とする。通常、MTUは静的または動的決定方法を使って決定される。いずれの方法でも、本発明は断片化を減らすことでネットワーク帯域幅を向上させることができる。一実施形態では、ネットワークのMTUは静的に決定される。この場合、データを送信する送信元ホストはMTUの値を合理的な程度に小さいMTU値、例えば576バイトに統計的に設定するか、これとは別に送信元ホストはMTUを送信元ホストとして同じローカル・ネットワークまたはサブネットに接続していないすべてのネットワークへのファーストホップMTUと576バイトのうち小さい方の値に設定する。

【0028】別の実施形態では、送信元ホストは所定の時間間隔でのインターネットの所与のルートに沿ってノードに問い合わせを行って、あるパスのMTUを動的に決定してパスMTU (PMTU) を決定する。PMTUは所与の特定のルートでの各ホップに関連付けられたMTUの最小値を示し、IPネットワークで送信されるデータグラムのMTUの上限値となる。インターネットでは、PMTUは「Path MTU Discovery」と呼ばれる技法を使って利用でき、Jeffrey MogulとSteve Deering著の、イン

ターネット特別技術調査委員会 (IETF) に関連して出版された「Path MTU Discovery」という題のRFC1191に記載されている。PMTUを実施する場合、本発明は、ネットワークでデータグラムが送信される際にホスト間でネットワーク状態およびローカル・ホスト・リソースに関する情報を交換できるホスト間通信機能が存在することをさらに前提としている。インターネットおよびその他のIPベースのネットワークで通常使われるホスト間データグラム通信の1つのタイプにインターネット制御メッセージ・プロトコル (ICMP) があり、J. Postel 著のIETFに関連して出版された「Internet Control Message Protocol」という題のRFC792に記載されている。Path MTU DiscoveryおよびICMPに関する上記の両文書を本発明の一実施形態で用いている。ICMPおよびPMTU Discoveryに類似したネットワーク・サービスを有するネットワーク上での作業に本発明の代替実施形態を適用できることを当業者は理解するであろう。

【0029】図6について説明する。同図は本発明の一実施形態によるデータグラム・パケットの最適サイズを決定するためのステップのすべてを含む流れ図である。まず、ステップ602で、本方法はインターネット上の送信元ホストと宛先ホストの間の所与のルートの最大伝送単位 (MTU) を決定する。前述したように、MTUの選択は時に困難な処理である。大半の接続レス・ネットワークでは、MTU値はルーティング・トポロジの変化に伴って急速に変動し、データグラムは異なるルートで宛先ホストに送信される。一次ルートが話中または物理的に切断されて使用不可になると、ネットワーク・トポロジが変化してパケットは別の二次ルートへルーティングされなくてはならない。通常、ルートはそれぞれ別々のMTU値を備えている。一実施形態では、上記のPMTU検出技法はデータ送信を行うたびにそれに先立って一度実行され、MTUが決定される。本発明により、このMTU値をデータグラム生成開始時点に使うことができる。また、本発明の別の実施形態では、データ送信中に所定の時間間隔で一連のPMTU検出を実行して、時が経つと発生する可能性があるMTUサイズの変動に対処することができる。一般に、これらのPMTU検出の周期はネットワークのパフォーマンス全体に影響が出るほど短くしてはならない。上記のRFC1191 Path MTU DiscoveryおよびRFC792 ICMPは、ICMPおよびPath MTU Discoveryを併用して所与のルートのMTUを決定する方法を簡潔に述べている。本発明の実施形態でも上記の静的MTU決定技法を使って所与のネットワークのMTU値を決定できるのはもちろんである。

【0030】次に、判定ステップ604で本方法は現在のデータグラムのサイズを上記のように決定したMTU

値と比較する。現在のデータグラムのサイズがMTU以下であれば、データグラムはそれより小さいデータグラムに断片化しなくても送信元ホストから宛先ホストへ送信できる程度に小さい。その場合、ステップ608でさらにデータグラム処理を行うためにデータグラムが送信される。しかし、データグラムのサイズがMTU値より大きい場合は、以下のステップでデータグラムをその後の断片化の確率が最小化される一連のより小さいデータグラムに分割する。

【0031】図6のステップ610で、MTU値と元のデータグラムまたはデータ・パケットの総バイト数を使っていくつのより小さいデータグラムを送信する必要があるかが決定される。ステップ610で、元のデータグラムに含まれる総バイト数が上で決定したMTU値で割られる。この除算の整数結果は一時可変DCOUNTに記憶され、余りがあれば一時可変RCOUNTに記憶される。これらの値は判定ステップ612で使われて元のデータグラムのデータを伝送するために追加のデータグラムが必要かどうか判定される。判定ステップ612で、RCOUNTに記憶された余りがゼロ値と比較される。比較の結果、RCOUNTが非ゼロの場合、処理はステップ614へ移行し、DCOUNT値を増分して追加のデータグラムが必要であることが示される。ただし、RCOUNT値がゼロの場合、元のデータグラム内のデータを送信するために追加のデータグラムは必要でなく、処理はDCOUNTを増分せずにステップ616へ移行する。

【0032】ステップ616で、元のデータグラム内のバイト数がDCOUNTに格納されている数だけのいくつかのデータグラムに均等に分配される。生成された小さいデータグラムはそれぞれ特定の断片に合わせてアドレス、オフセット、チェックサム、およびその他の特定の値が調整された状態で元のデータグラムに含まれる同じヘッダ情報を受信することを当業者は理解するであろう。従来技術によって生成されたデータグラムとは対照的に、これらのデータグラムはデータグラム値に大小のばらつきがない。その代わりに、バイトはより小さいデータグラム・セグメントのそれぞれに均等に分配される。こうしてできたデータグラムはサイズがほぼ等しく、MTUが小さいルートでデータグラムが送信される際に断片化される可能性はより低い。

【0033】例として、図1に送信元ノード102と宛先ノード114の間に2つのルートがあるネットワークを示す。ルートAはネットワーク104、108、および112を含み1100バイトのパスMTU (PMTU) を備え、ルートBはネットワーク104、118、および112を含み1006バイトのパスMTU (PMTU) を備える。図5Bにその種のネットワーク上で送信するために従来技術で1700バイトのデータグラムを2つのより小さいデータグラムに分割する方法を示

す。図5Bの第1のデータグラムはルートA(PMTU1100)を通る場合にルート106を通ることは可能だが、さらに断片化をしないとルート110経由では不可能である。同様にルートB(PMTU1006)では、第1のデータグラムはルート106からノード116を通るのに断片化が必要である。図5Bの第2のデータグラムはさらに断片化をしなくてもいずれのルートも通ることができる。さらなる断片化に関連付けられた追加のオーバーヘッドの結果、オーバーヘッドが発生してネットワーク効率は低下する。

【0034】対照的に、図5Dは本発明の各実施形態が同じ数のバイトを同じ数のデータグラムにより最適な方法で分割する様子を示す。いくつかのデータグラム内のバイト数を増やして他のデータグラムのバイト数を減らすのではなく、本発明は送信データを最小数のデータグラムに均一に分割する。この例では、図5Dの両方のデータグラムはわずか850バイトしか含んでいない。したがって、これらのデータグラムは両方とも断片化なしに図1のルートAとルートBを通過できる。

【0035】本発明の実施形態には当業界で以前は不可能だったいくつかの有利な点がある。第1に、本発明で提供された技法では送信されるデータグラム数は実質的に変わらないため、ネットワーク・オーバーヘッドの量を増やさずに断片化の確率が低下する。最良のケースではこれらの技法は大幅に断片化を減らして有効ネットワーク帯域幅を拡張するが、最悪のケースでもこれらの技法の有効性は他の技法と同じ程度である。

【0036】本発明の実施形態はネットワーク・プロトコルに非互換性を持ち込まずにネットワーク・パフォーマンスを向上させるため、また有利である。大半の大きいネットワークではネットワーク・プロトコル内の低レベル層の変更は不可避免的に非互換となりネットワークの相互運用性が減少する。これは特にIPやTCP/IPなどのプロトコルを実行する数百、数千のコンピュータが直ちに變更できないインターネットについて当てはまる。本明細書に記載したIP層の改良は基本的にIPプロトコル動作を變更しないため、本発明を使用してもこのジレンマは発生しない。したがって、本発明の教示によってネットワークの相互運用性を維持し、有効ネットワーク帯域幅を改善するという困難な問題は明快に解決される。

【0037】以上、本明細書で特定の実施形態について例示してきたが、本発明の精神と範囲を逸脱することなしにさまざまな修正を加えることができる。本発明の実施形態は多様な異なるネットワーク・プロトコルを使って実施でき、TCP/IPプロトコルを使用するネットワークに接続したコンピュータ・システムに限定されないことを当業者は理解するであろう。好ましい実施形態に実質的に類似した代替実施形態をSNA(システム・ネットワーク・アーキテクチャ)、IPX/SPX、A

ppletalk、またはX.25などのその他のネットワーク・プロトコルを使って実施できる。TCP/IPおよびSNAネットワークの統合に関する詳細については、Taylor著「Integrating TCP/IP into SNA」、Wordware Publishers、1993年を参照されたい。さらに、好ましい実施形態に実質的に類似した別の代替実施形態をPMTU値が一回ただ収集されるだけではなく所定の時間間隔で無数の回数更新されるという点を除いて実施できる。好ましい実施形態に実質的に類似したさらに別の代替実施形態を最大パケット・サイズ送信制限があるすべてのネットワークに適用できるという点を除いて実施できる。本発明の例では一般にバイト数を使ってデータグラムのサイズを測定しているが、データグラムのサイズはいかなる他の伝送単位の尺度、例えば、時間間隔が測定するビット数、ブロック数、またはデータ・ストリーム数によってさえ測定できることを当業者は理解するであろう。

【0038】これらの実施形態は、コンピュータを使用するものであるので、当然ソフトウェアを含み、それらはCDROMその他のコンピュータが読むことのできる記録媒体にコードとして具体化して記録される。その媒体に記録されるコードは、それぞれのルートが異なるサイズのパケットを伝送できるネットワーク上の異なるルートで送信されるパケットを生成できるように記録されている。さらに、そのコードは、ネットワーク上で送信されるパケット内の記憶単位の総数をネットワーク上の所与のルートに関連する最大伝送単位(MTU)と比較するように構成された第1のコード部分と、比較の結果がパケット内の記憶単位の総数がMTUより大きいことを示している場合にパケットをさらに処理する第2のコード部分と、パケット内の記憶単位数をMTU記憶単位を保持できる1つまたは複数のパケット内に記憶された実質的に均等な単位のグループに分離するように構成された第3のコード部分とを備えている。

【0039】以上のように、本発明は上述した実施形態に限定されず、むしろ同じ価値のある物の完全な範囲にわたって添付の特許請求の範囲によって定義される。

【図面の簡単な説明】

【図1】 本発明の一実施形態を実施するコンピュータ・ネットワークを示す図である。

【図2】 ネットワーク上のルータまたはノードとして動作するように構成されたコンピュータ・システムの必須コンポーネントを示すブロック図である。

【図3】 コンピュータ・ネットワーク上の例示的なルータ装置およびノードで用いられる異なるプロトコル層を示すブロック図である。

【図4】 IP層を使ってパケット交換ネットワークの送信元ノードから宛先ノードへデータを送信する方法を示すブロック図(A)とIPネットワークで送信される

データ・パケットの代表的な変更を示す図（B）である。

【図5】 代表的な従来技術の方法が大きいデータグラムをいくつかのデータグラムに不均等に断片化する方法を示す図（A、B）と本発明の一実施形態を使って大きいデータグラムをいくつかのデータグラムに均等に断片化する方法を示す図（C、D）である。

【図6】 本発明の一実施形態にしたがってデータグラム・パケットを生成するために用いられる全ステップを示す流れ図である。

【符号の説明】

202 コンピュータ・システム

211 第2のネットワーク・インタフェース

212 第1のネットワーク・インタフェース

214 プロセッサ

216 一次記憶

218 二次記憶

220 入出力インタフェース

222 オペレーティング・システム

224 プロトコル・スタック

226 データグラム生成器

302 ノード

306 ルータ装置

310 ノード

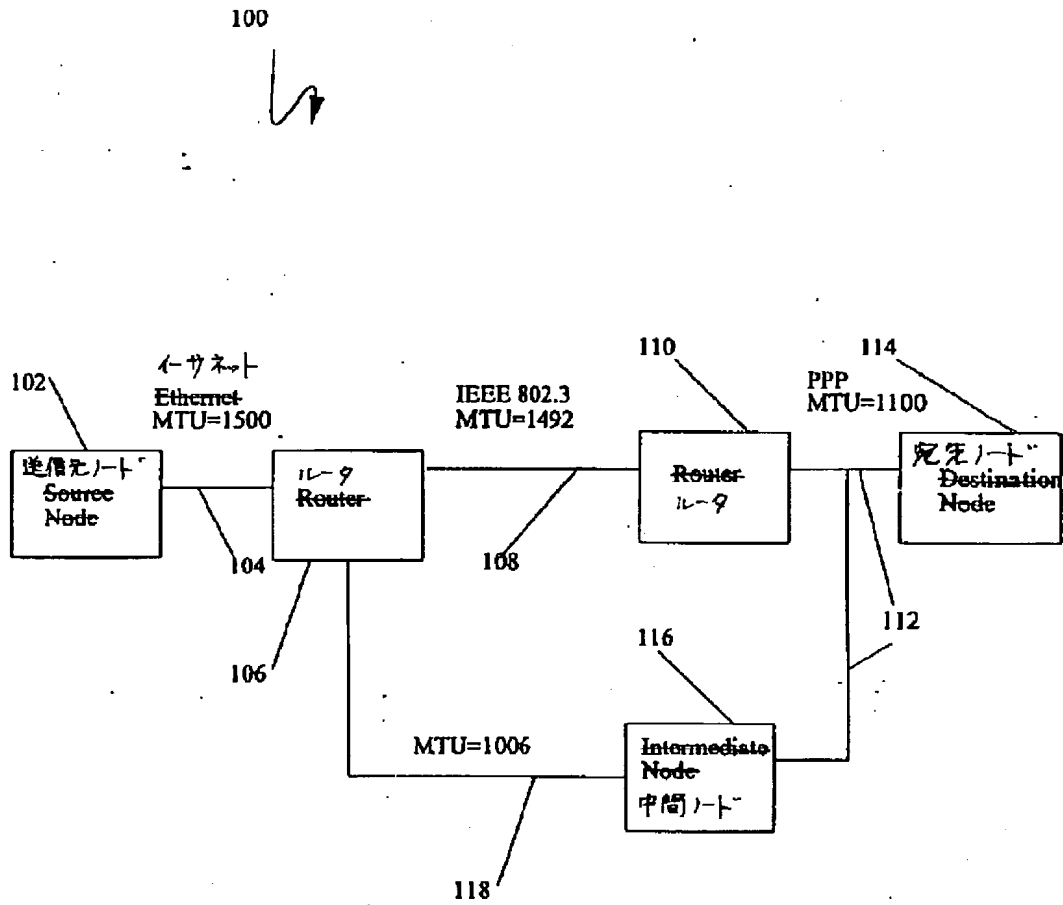
402 第1のノード

407 X.25パケット交換ネットワーク

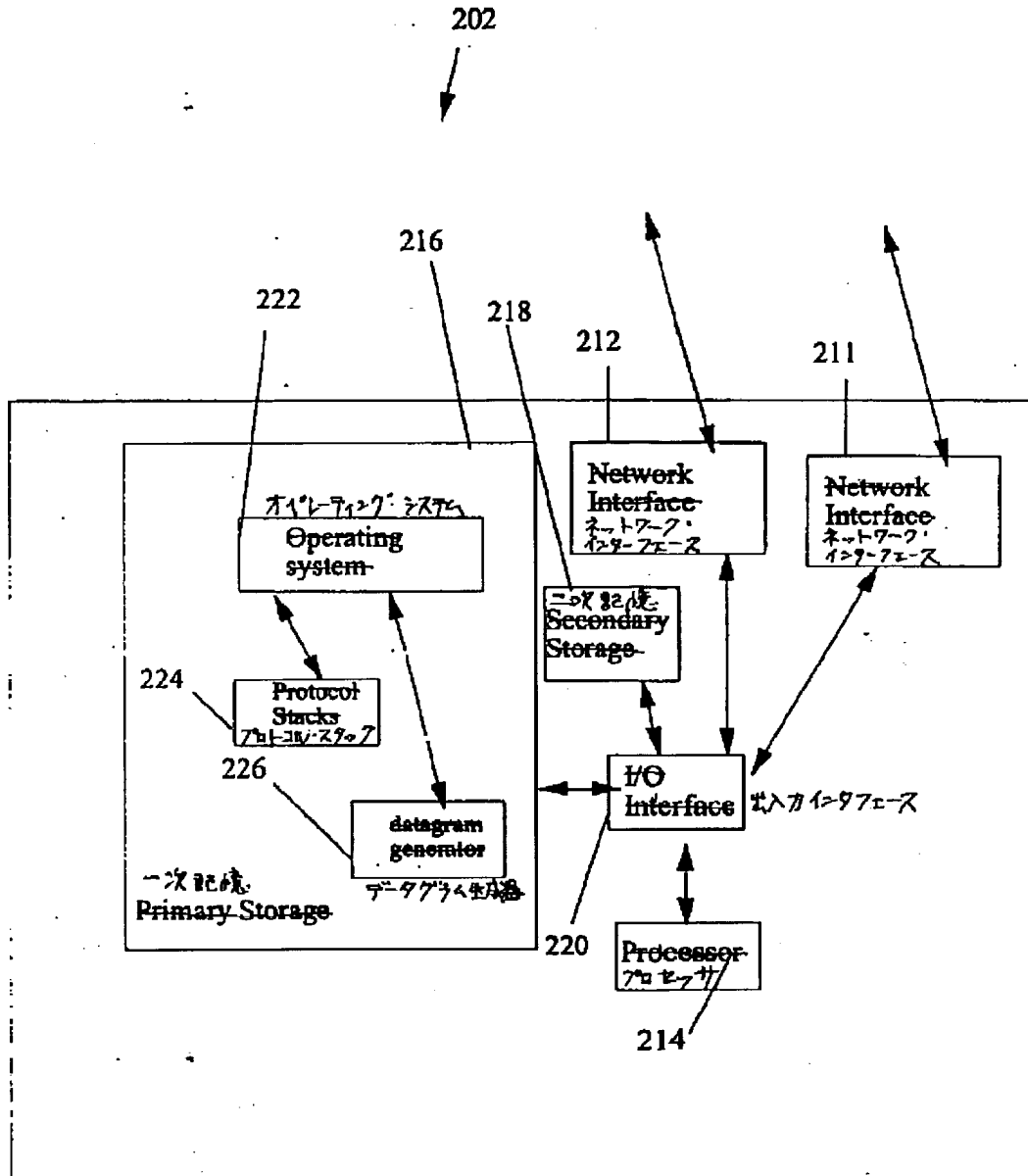
410 ルータ

412 イーサネット・ローカル・エリア・ネットワーク

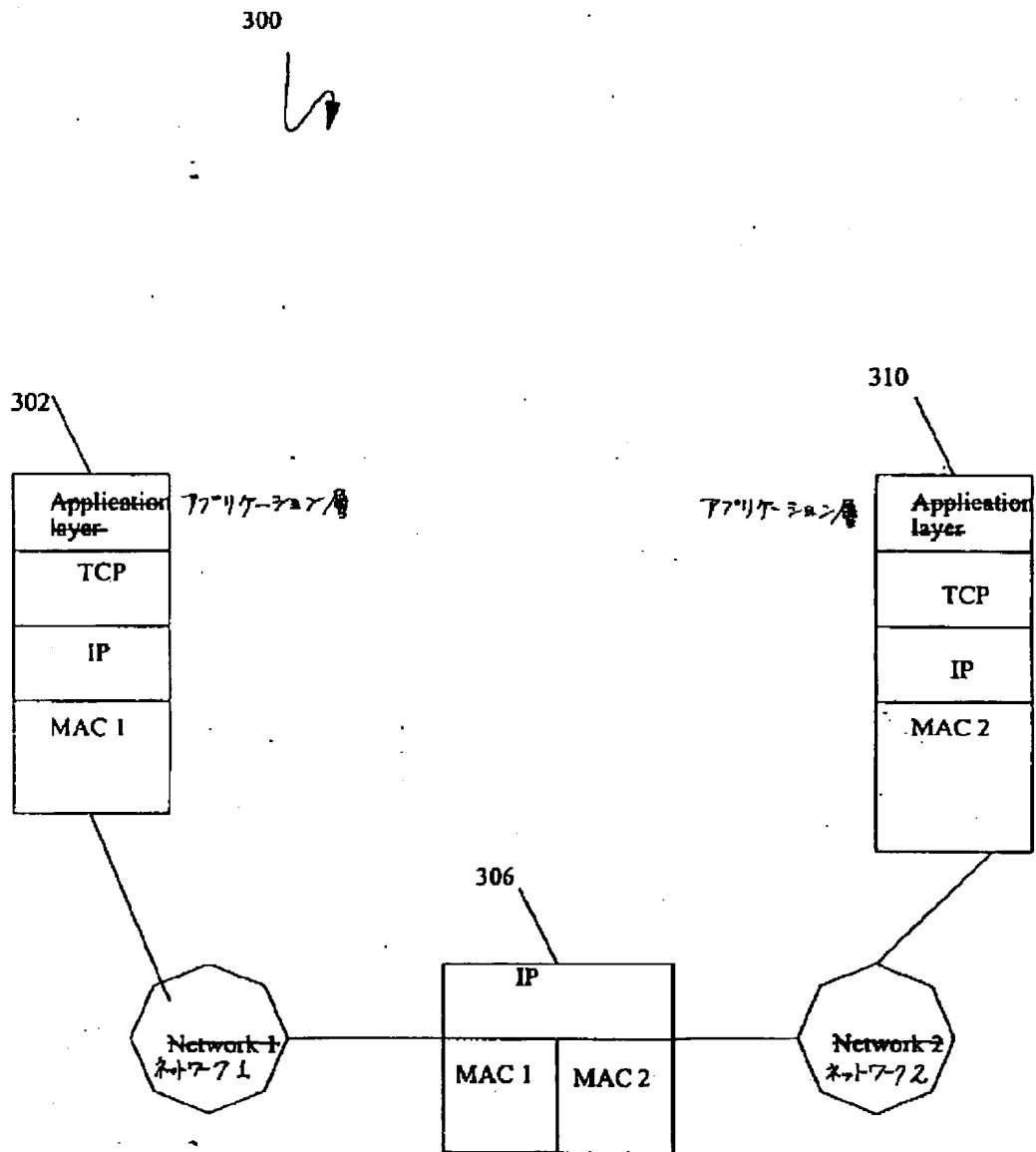
【図1】



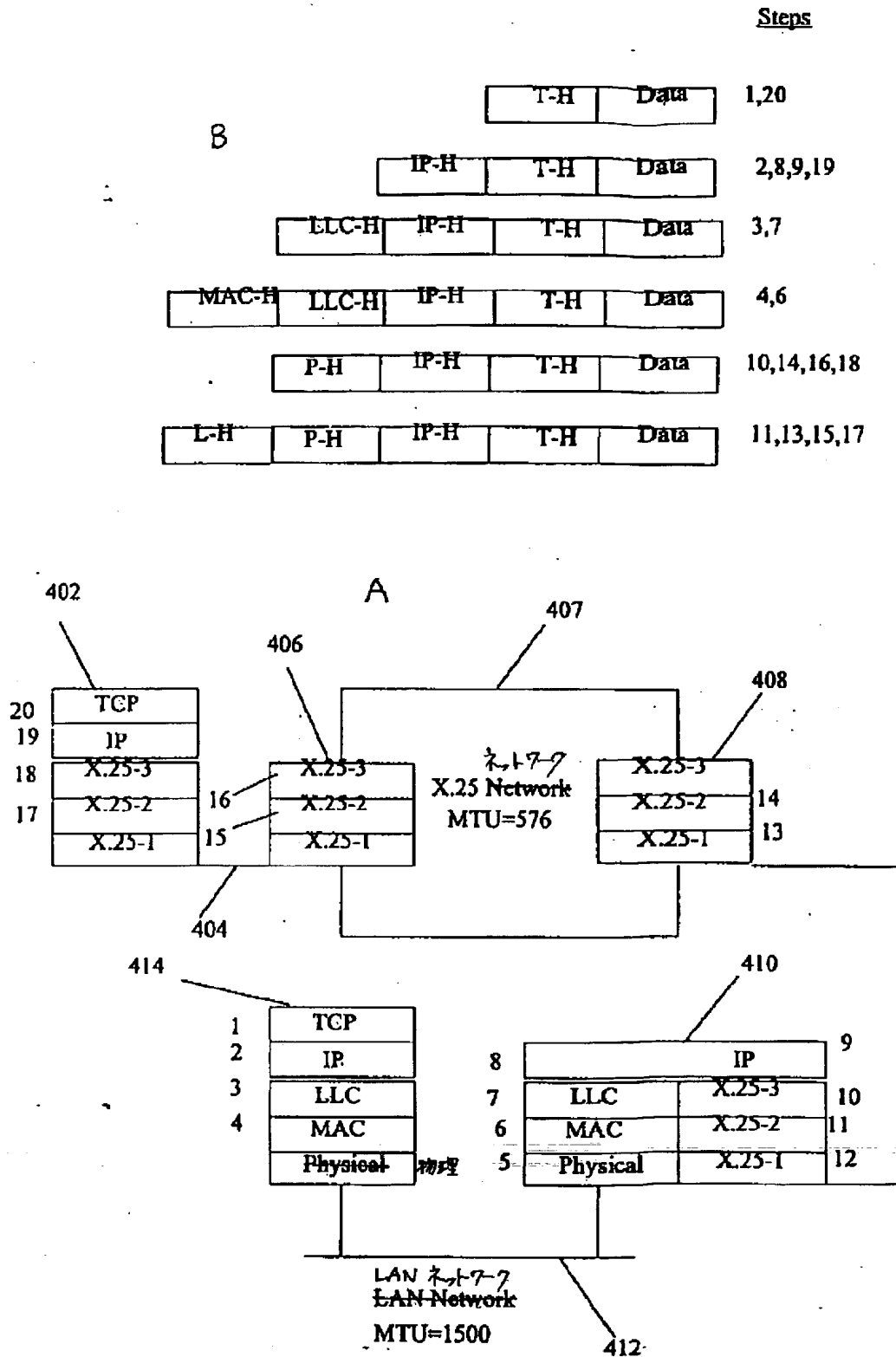
【図2】



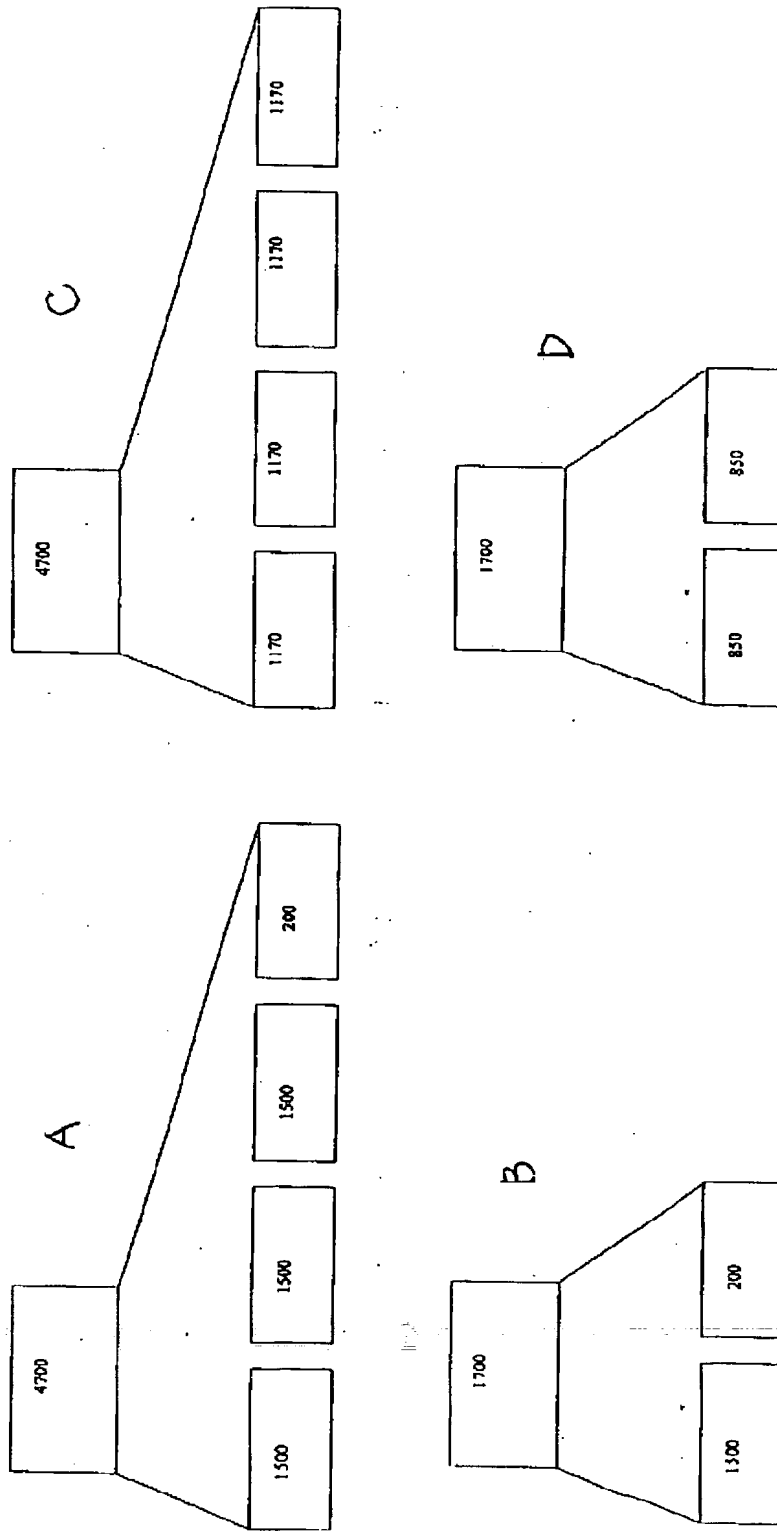
【図3】



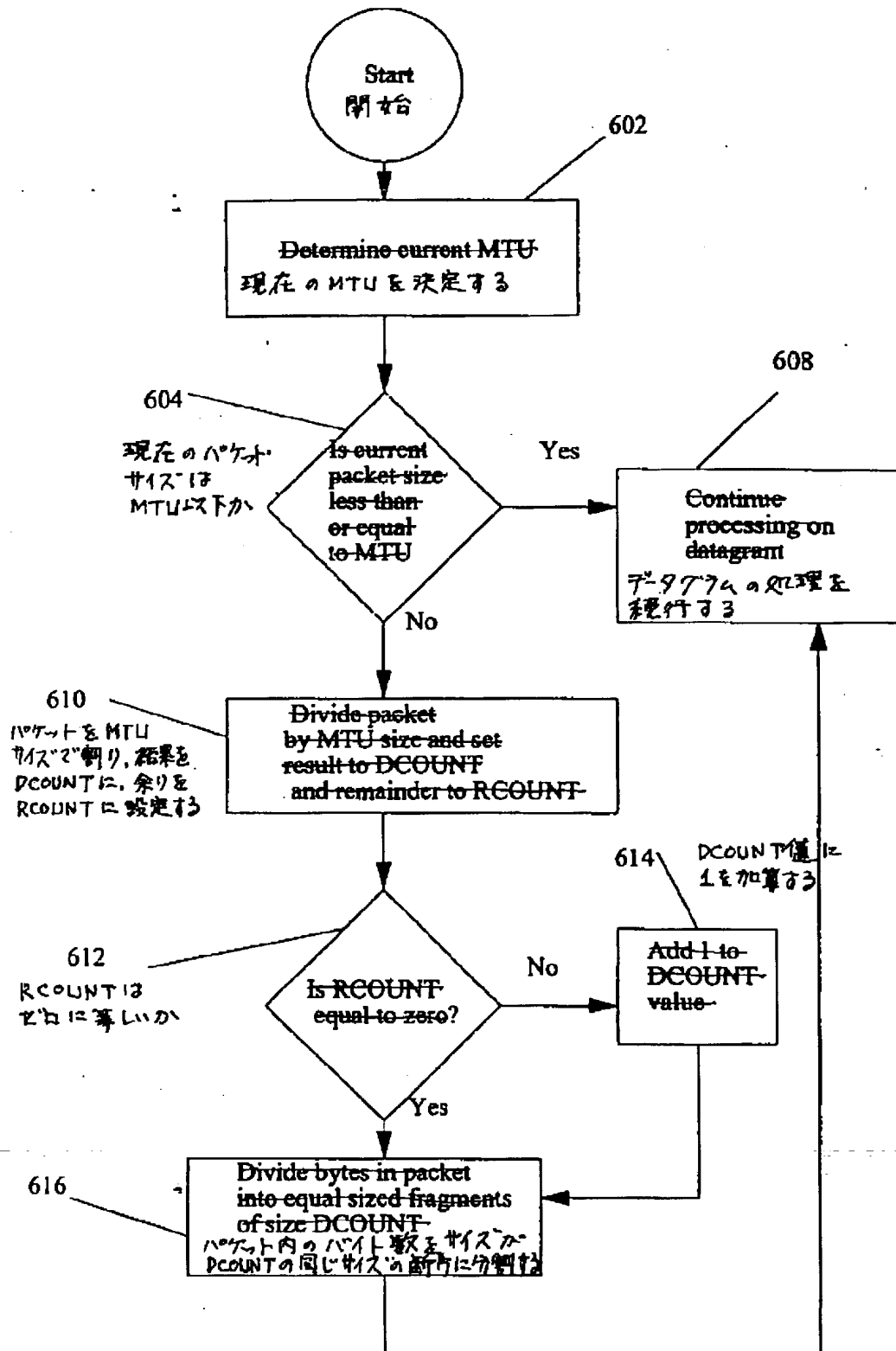
【図4】



【図5】



【図6】



【手続補正書】

【提出日】平成10年8月5日

【手続補正1】

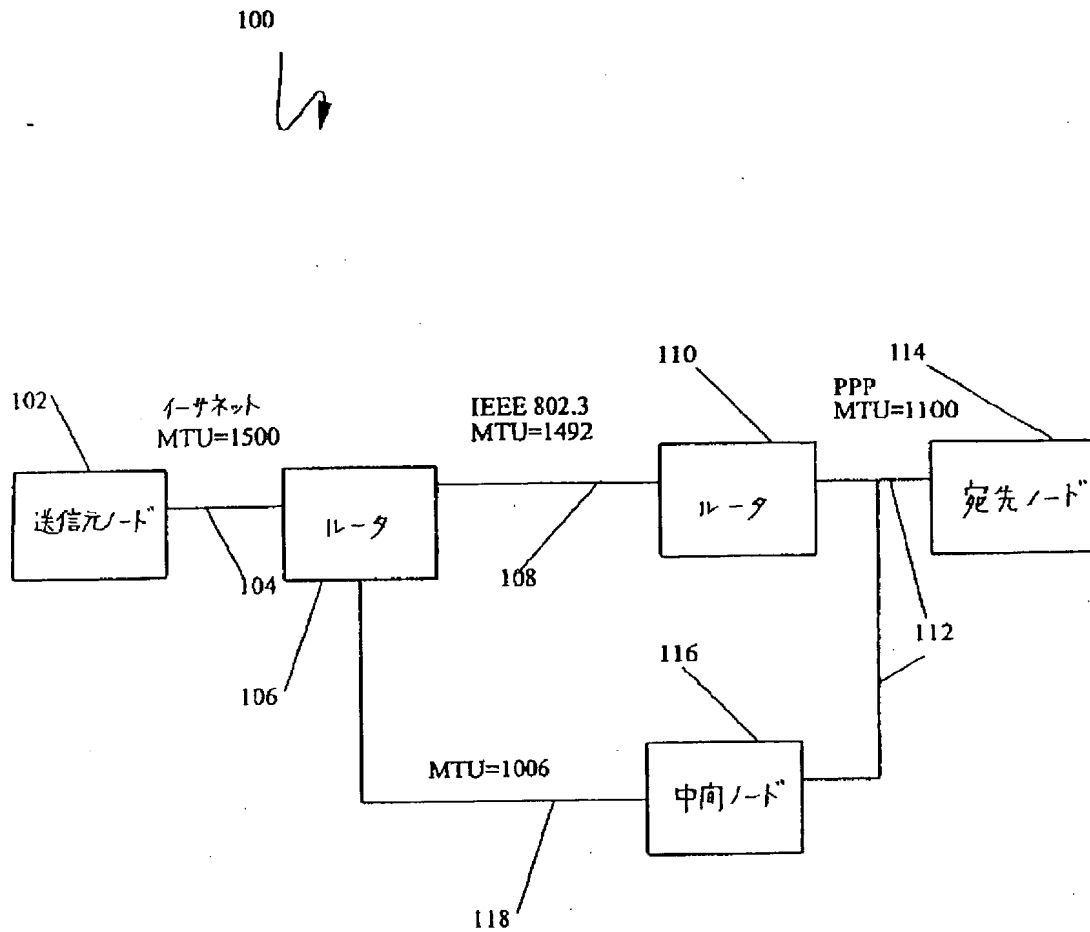
【補正対象書類名】図面

【補正対象項目名】全図

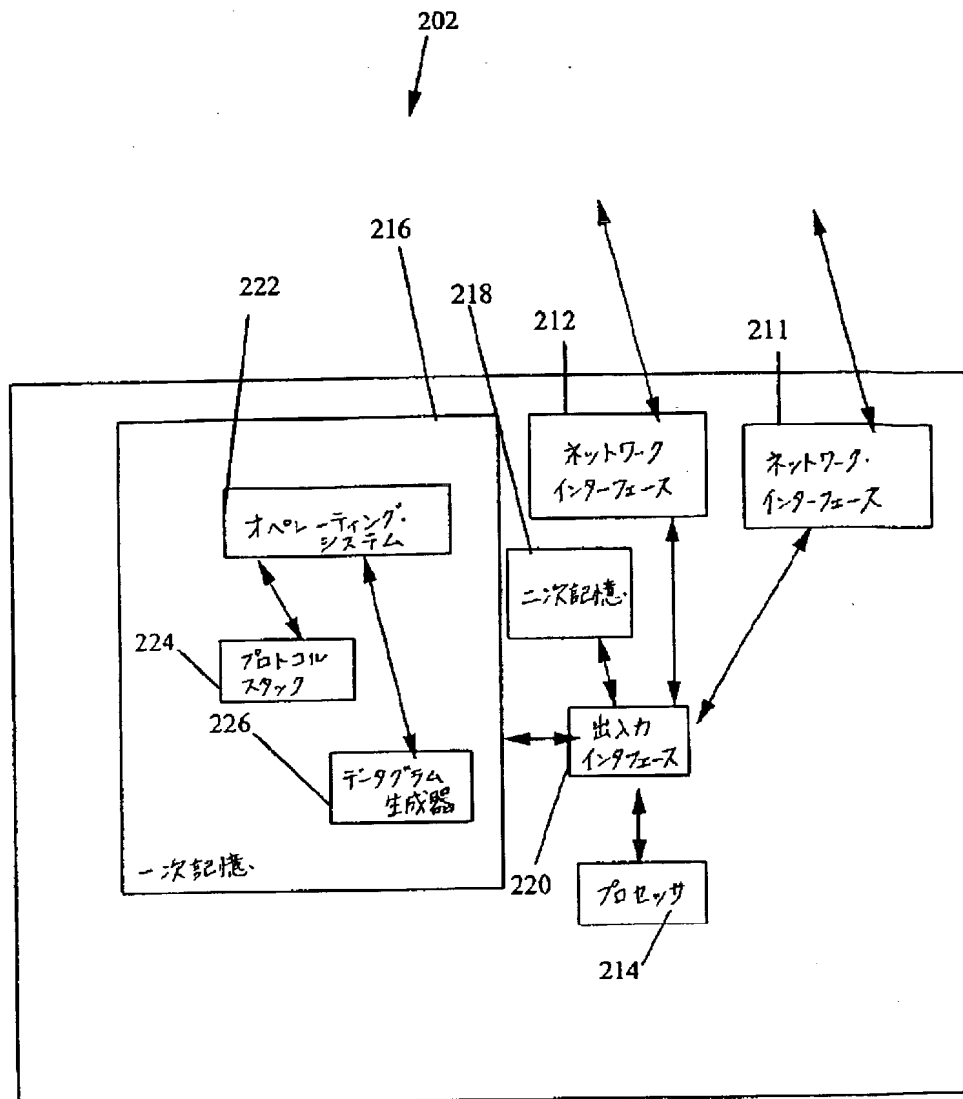
【補正方法】変更

【補正内容】

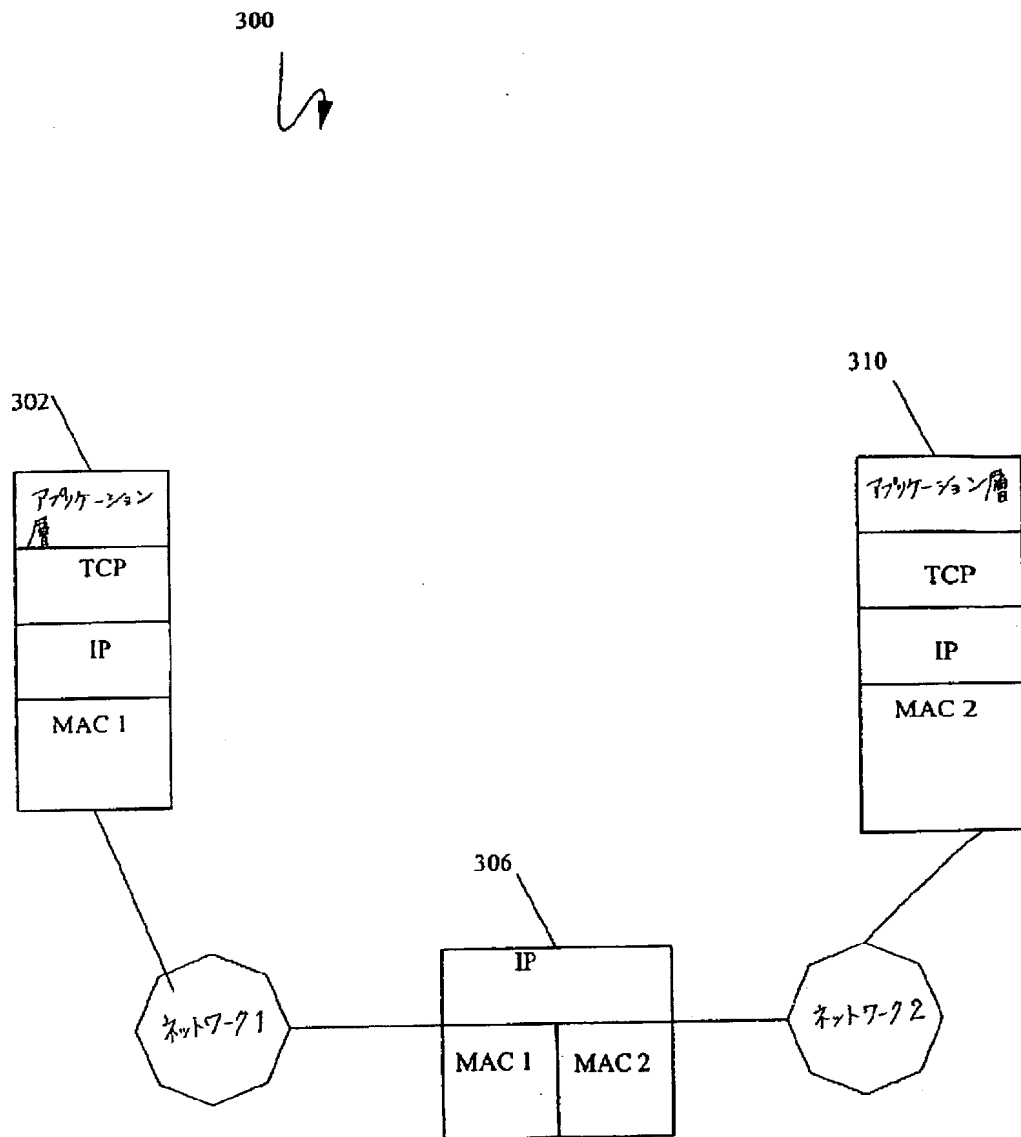
【図1】



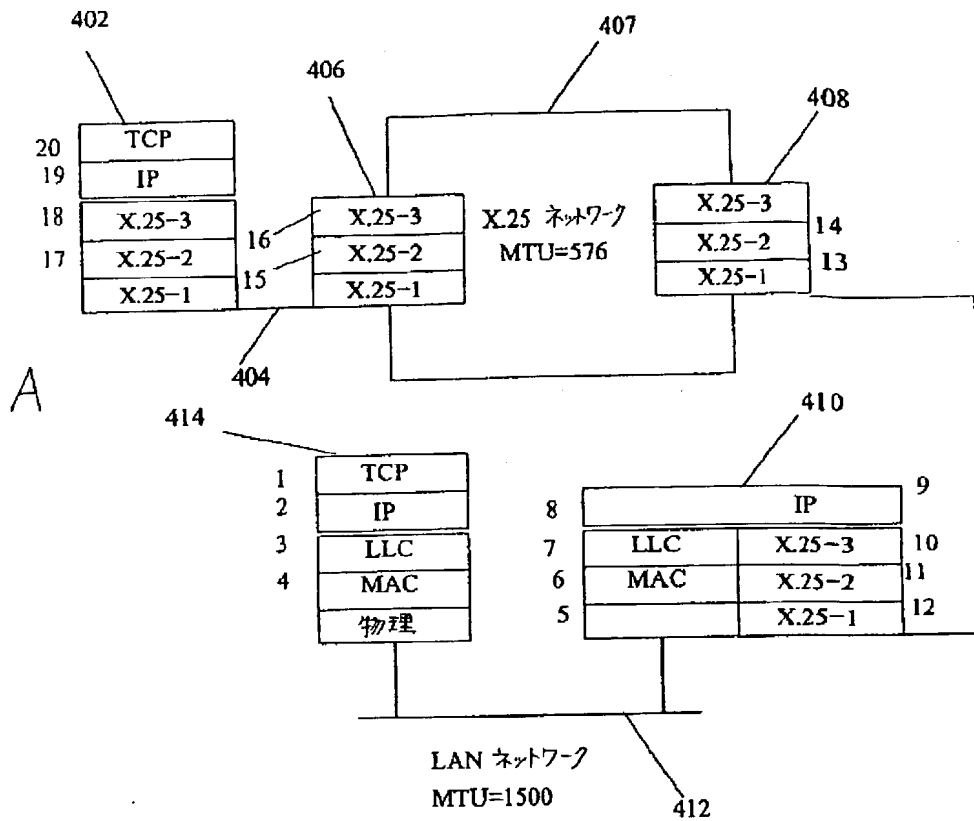
【図2】



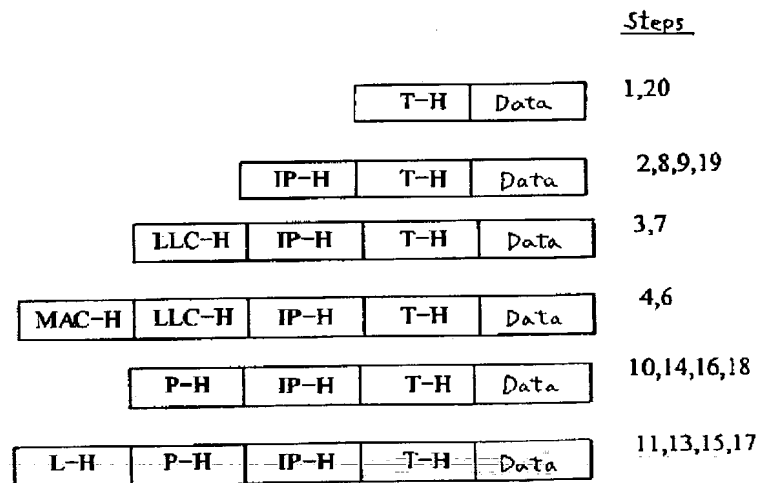
【図3】



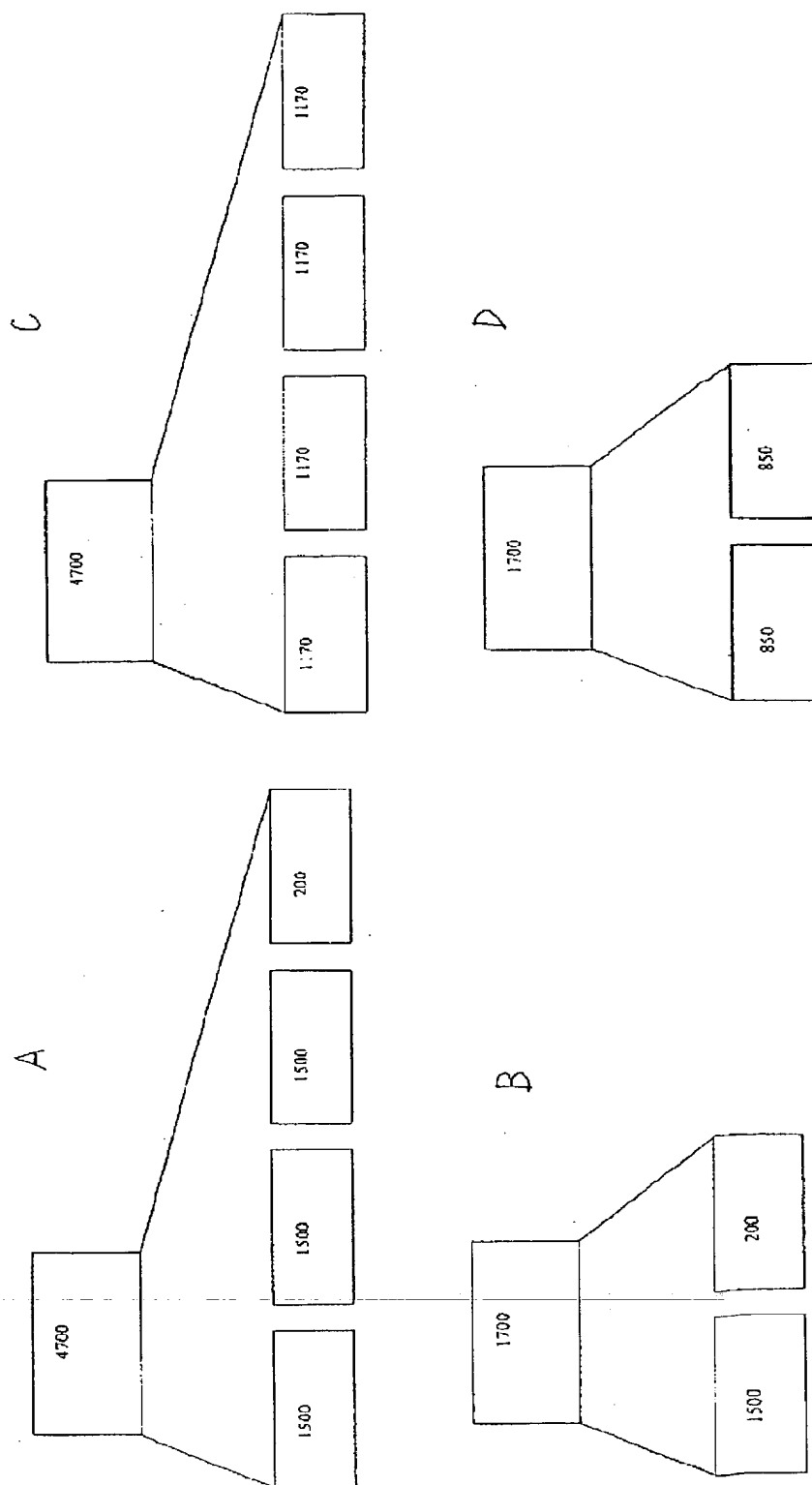
【図4】



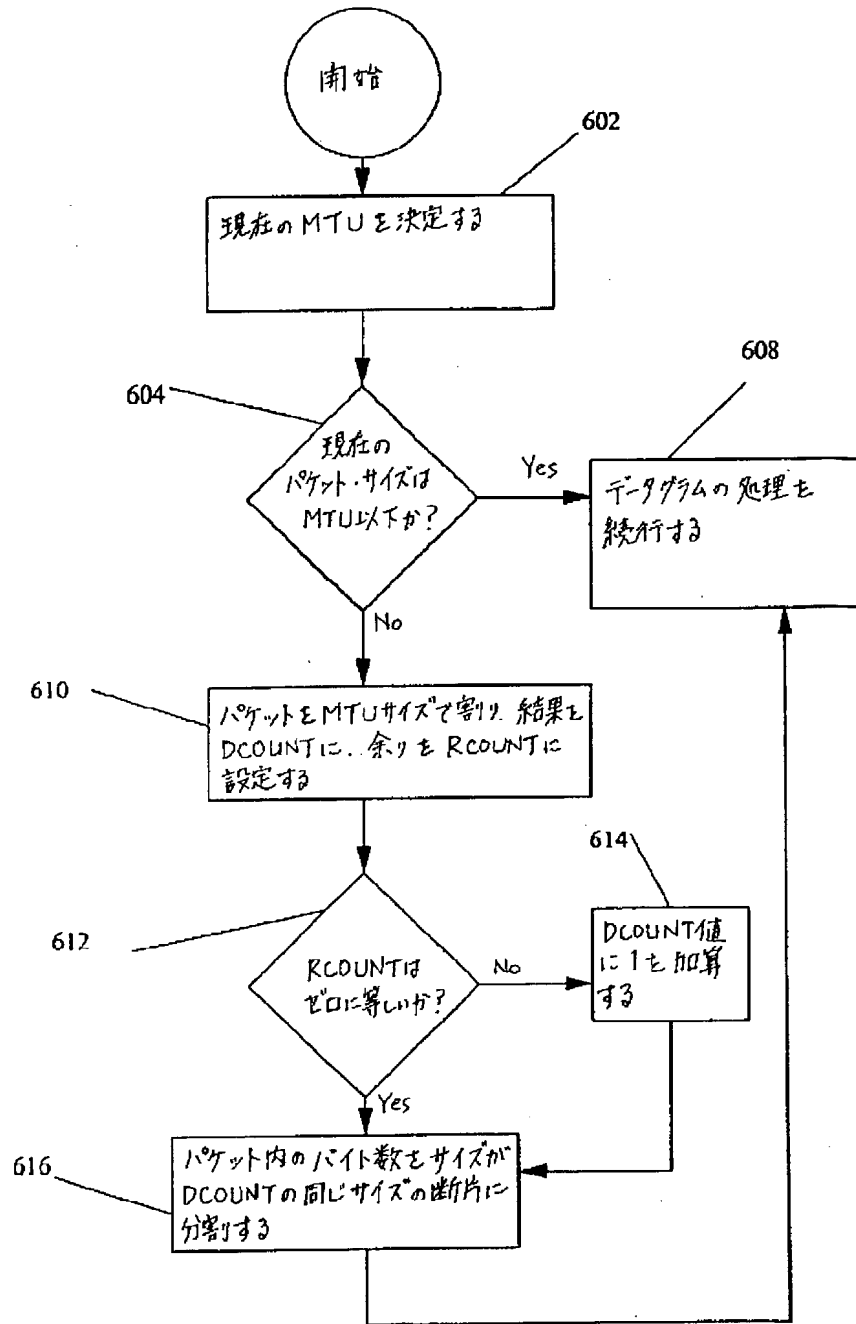
B



【図5】



【図6】



フロントページの続き

(71)出願人 591064003
901 SAN ANTONIO ROAD
PALO ALTO, CA 94303, U.
S. A.